Efficient MinHash-based Algorithms for Big Structured Data

Wei Wu

Faculty of Engineering and Information Technology

University of Technology Sydney

A thesis is submitted for the degree of

Doctor of Philosophy

July 2018

I would like to dedicate this thesis to my loving parents ...

Declaration

I hereby declare that except where specific reference is made to the work of others, the contents of this dissertation are original and have not been submitted in whole or in part for consideration for any other degree or qualification in this, or any other University. This dissertation is the result of my own work and includes nothing which is the outcome of work done in collaboration, except where specifically indicated in the text. This dissertation contains less than 65,000 words including appendices, bibliography, footnotes, tables and equations and has less than 150 figures.

Wei Wu
July 2018

Acknowledgements

I would like to express my earnest thanks to my principal supervisor, Dr. Ling Chen, and co-supervisor, Dr. Bin Li, who have provided the tremendous support and guidance for my research. Dr. Ling Chen always controlled the general direction of my PhD life. She made the recommendation letter for me, and contacted the world leading researchers for collaboration with me. Also, she encouraged me to attend the research conferences, where my horizon was remarkably broaden and I knew more about the research frontiers by discussing with the worldwide researchers. Dr. Bin Li guided me to explore the unknown, how to think and how to write. He once said to me, "It is not that I am supervising you. Instead, this is a spiraling process and we make progress together." In the first two years of PhD, I was still groping in my research topic and did not obtain any results. During that time, I was very anxious and unconfident. However, my supervisors were patient enough and always encouraged me. No doubt that I cannot acquire my current research performance without their patient supervision.

I am grateful to my best friend, Dr. Chuan Luo, who shares three-year research life in Peking University. Although we are currently in different research communities, we often discuss the research trends. In my most depressing days, he enlightened me patiently, even though we were in different countries.

I am grateful to all members in CAI for their participation in and valuable comments on my research. Particularly, I thank Prof. Chengqi Zhang, Dr. Chunyang Liu, Dr. Sujuan Hou, Dr. Meng Fang, Dr. Qinzhe Zhang, Dr. Weiwei Liu, Mr. Yu Liu, Mr. Bo Han, Mr. Yuangang Pan, Mr. Jiangchao Yao, Mr. Adi Lin, Mr. Daokun Zhang, Mr. Dong Wen, Dr. Qian Zhang and Miss Jiamiao Wang, and other students for their help in my PhD study.

I am grateful to the financial support from UTS scholarship and FEIT Travel funds.

Finally, and above all, I would like to thank my parents for their unconditional support and encouragement from emotion and finance. They always believe in me, give me invaluable suggestions and fully support all my decisions. No words could possibly express my deepest gratitude for their endless love, self-sacrifice and unwavering help.

To them I dedicate the dissertation.

Abstract

Nowadays, data are growing explosively in the information society. Every day, an incredibly large number of data are generated from social networks, financial industries, medical device and digital equipment, etc. – Big data have been driving data mining research in both academia and industry. No matter how data mining and machine learning develop, in most tasks such as classification, clustering and retrieval, it is one of the most fundamental operations to compute the similarity of data. However, exact similarity computation has become daunting for big data due to the "3V" nature (volume, velocity and variety). Therefore, the thesis aims to develop a number of efficient randomized hashing algorithms for big structured data to approximate the similarity.

Taking into account the popularity and importance of data modelled as weighted sets, graphs/networks, the MinHash-based algorithms are proposed to transform weighted sets, graphs and network nodes into low-dimensional hash codes for efficient similarity estimation.

In Chapter 4, we propose two MinHash-based algorithms for weighted sets, which are two instances of the Consistent Weighted Sampling (CWS) scheme, for real-value weighted sets by uncovering the working mechanism of the state-of-the-art weighted MinHash algorithm, Improved Consistent Weighted Sampling (ICWS) algorithm. The two algorithms both represent a weighted set as a low-dimensional hash code. The first algorithm, the Canonical Consistent Weighted Sampling (CCWS) algorithm, reconstructs the hash functions in ICWS in order to overcome the flaw that ICWS breaks the uniformity of the MinHash scheme. Consequently, our CCWS algorithm not only shares the same theoretical computational complexity as ICWS but also strictly complies with the definition of the MinHash algorithm. The second algorithm, the Practical Consistent Weighted Sampling (PCWS) algorithm, simplifies ICWS by transforming the original form of ICWS into an equivalent expression. As a result, our PCWS algorithm is not only mathematically equivalent to ICWS and preserves the same theoretical properties, but also saves 20% memory footprint and substantial computational cost compared to ICWS.

In Chapter 5, we propose a MinHash-based algorithm for graphs, i.e., the K-Ary Tree Hashing (KATH) algorithm, for representing a graph as a low-dimensional feature vector.

The main idea of KATH is to construct a traversal table to quickly approximate the subtree patterns in Weisfeiler-Lehman (WL) graph kernel using K-ary trees. Based on the traversal table, our KATH algorithm employs a recursive indexing process that performs only r times of matrix indexing to generate all $(r-1)$-depth K-ary trees, where the leaf node labels of a tree can uniquely specify the pattern. After that, the MinHash scheme is used to fingerprint the acquired subtree patterns for a graph. The experimental results show that our KATH algorithm runs significantly faster than state-of-the-art methods while achieving competitive or better accuracy.

In Chapter 6, we propose a MinHash-based algorithm for network nodes, i.e., the NetHash algorithm. Consequently, each node in the network is mapped to a low-dimensional vector space. Our NetHash algorithm employs the randomized hashing technique to encode shallow trees, each of which is rooted at a node of the network. The main idea is to efficiently encode both attributes and structure information of each node by recursively sketching the corresponding rooted tree from bottom (i.e., highest-order neighboring nodes) to top (i.e., root node), and particularly, to preserve as much information closer to the root node as possible. The extensive experimental results show that the proposed algorithm, which does not need learning, runs significantly faster than the state-of-the-art learning-based network embedding methods while achieving competitive or even better performance in accuracy. Furthermore, the simplicity of the algorithm enables our NetHash algorithm to embed a millions-of-nodes network within 1,000 seconds on a single machine.

Table of contents

List of figures

List of tables